ST. LAWRENCE HIGH SCHOOL

A JESUIT CHRISTIAN MINORITY INSTITUTION

FOR GOD AND COUNTRY

STUDY MATERIAL-2
SUBJECT – STATISTICS

Pre-test

Chapter: BIVARIATE ANALYSIS **Class: XII**

Topic:CORRELATION **Date: 11.05.2020**

CORRELATION

PART 2

Some important problems on correlation

Q1 If x_1, x_2 and x_3 are uncorrelated variables each having same standard deviation, obtain the correlation coefficient between $(x_1 + x_2) and (x_2 + x_3)$.

Ans: Let v $(x_1) = v(x_2) = v(x_3) = k^2$

Cov $(x_1 + x_2, x_2 + x_3) = cov(x_1, x_2) + cov(x_1, x_3) + cov(x_2, x_3) + v(x_2)$

$= 0 + 0 + 0 + k^2 = k^2$

$V(x_1 + x_2) = v(x_1) + v(x_2) + 2\,cov(x_1, x_2) = 2k^2$

Similarly $V(x_3 + x_2) = v(x_3) + v(x_2) + 2\,cov(x_3, x_2) = 2k^2$

Hence $r_{x_1+x_2, x_2+x_3} = \frac{\text{Cov}(x_1+x_2, x_2+x_3)}{\sqrt{V(x_1+x_2)V(x_3+x_2)}} \frac{k^2}{2k^2} = \frac{1}{2}$.

Q2. x and y are two variables with standard deviations $s_x and s_y$, respectively. They have positive correlation r. Determine the value of the constant k such that $(x+ky) and \left(x + \frac{s_x}{s_y}.y\right)$ are uncorrelated.

Ans: According to the question,

Cov $\left(x + ky, x + \frac{s_x}{s_y} y\right) = 0$

$\Rightarrow s_x^2 + k \frac{s_x}{s_y} s_y^2 + \left(k + \frac{s_x}{s_y}\right) cov(x, y) = 0$

$\Rightarrow s_x^2 + k\, s_x s_y + \left(k + \frac{s_x}{s_y}\right) s_x s_y . r = 0$

$\Rightarrow k = \frac{-s_x^2 (1+r)}{s_x s_y (1+r)} = \frac{-s_x}{s_y}$.

Q3. Two variables x and y take the values as follows :

x	:	–3	–1	+1	+3
y	:	9	1	1	9

Show that their correlation coefficient is zero. Are the variables independent? If not, what is the reason for correlation coefficient being zero?

Ans: $\sum x_i = 0 \; and \; \sum x_i y_i = 0 \implies cov\,(x, y) = 0$ but from the given data $y = x^2$.

Correlation coefficient is zero implies that x and y are linearly independent but here these are not linearly independent.

Q4. Obtain correlation coefficient r for the accompanying data on y = glucose concentration (g/L) and x = fermention time (in days) for a particular brand of a malt liquor :

x	:	1	2	3	4	5	6	7	8
y	:	74	54	52	51	52	53	58	71

Interpret the value of r you obtain.

What is your impression about the relationship between glucose concentration and fermention time?

Ans: take $u_i = x_i - 4 \; and \; v_i = y_i - 52$

$\sum u_i = 4 \; and \; \sum v_i = 49 \; and \; \sum u_i v_i = 26$

$\sum u_i^2 = 44 \,, \; \sum v_i^2 = 887$

Cov (u, v) = $\frac{1}{8} \,, \sum u_i v_i - \bar{u}\bar{v} = \frac{26}{8} - \frac{4}{8}\,\frac{49}{8} = 0.1875$

$s_u^2 = \frac{1}{8}\,.44 - (\frac{4}{8})^2 = 5.4375 \implies s_u = 2.33$

$s_v^2 = \frac{1}{8}\,887 - (\frac{49}{8})^2 = 73.36 \implies s_v = 8.56$

$r_{uv} = 0.009 \implies r_{xy} = 0.009$. So the variables are positively correlated but not in a high extent. Since the value is very close to 0, those are almost linearly independent.

Q5. While calculating the coefficient of correlation between variables x and y the following results were obtained :

$n = 25, \sum x = 125, \sum y = 100, \sum x^2 = 650, \sum y^2 = 460$, $\sum xy = 508$. It was however later discovered at the time of checking that 2 pairs of observations (x,y) were copied wrongly as (6,14) and (8,6) while the correct values were (8,12) and (6,8) respectively.

Determine the correct correlation coefficient.

Ans: Corrected values

$\sum x = 125,\ \sum y = 100\ ,\ \sum x_i^2 = 650\ ,\sum y_i^2 = 436,\ \sum x_i y_i = 520$

Cov(x, y) = $\frac{520}{25} - \frac{125}{25} X \frac{100}{25} = 0.8,\ s_x = \sqrt{\frac{650}{25} - (\frac{125}{25})^2} = 1$

$s_y = \sqrt{\frac{436}{25} - (\frac{100}{25})^2} = = 1.44.\ So\ r_{xy} = 0.56$

Q6.. If x' *and* y' are the deviations of variables x and y from their means, s_1^2, s_2^2 are the variances of x and y, correlation coefficient between x and y is r, and (x_1, y_1) *for* $i = 1,2,...,n$ are the n values of (x,y), show that

$$r = 1 - \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{x_i'}{s_1} - \frac{y_i'}{s_2}\right)^2 = -1 + \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{x_i'}{s_1} + \frac{y_i'}{s_2}\right)^2$$

Hence prove that $-1 \le r \le 1$.

Ans: $\sum x_i'^2 = \sum(x_i - \bar{x})^2 = n\, s_x^2$ and similarly $\sum y_i'^2 = \sum(y - \bar{y})^2 = n\, s_y^2$

$\sum x_i' y_i' = \sum\big((x_i - \bar{x})(y_i - \bar{y})\big) = n\, cov(x, y)$

$\frac{1}{2n}\sum(\frac{x_i'}{s_1} \pm \frac{y_i'}{s_2})^2 = \frac{1}{2n}\ (\sum\frac{x_i'^2}{s_1^2} + \sum\frac{y_i'^2}{s_2^2}\ \pm \sum 2\frac{x_i'}{s_1}\ \frac{y_i'}{s_2}) = \frac{1}{2n}(\ n + n\ \pm 2nr)$

$= 1 \pm r$

Hence $r = 1 - \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{x_i'}{s_1} - \frac{y_i'}{s_2}\right)^2 = -1 + \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{x_i'}{s_1} + \frac{y_i'}{s_2}\right)^2$

Since square quantities are always greater than zero , so $-1 \le r \le 1$.

Q7. Two variables x and y with respective standard deviations s_x *and* s_y, have correlation coefficient r. The variable u and v are defined as

$$u = x\cos\alpha + y\sin\alpha$$

$$v = y\cos\alpha - x\sin\alpha$$

Show that

(i) u and v will be uncorrelated if

$$\tan 2\alpha = \frac{2rs_x s_y}{s_x^2 - s_y^2}$$

(ii) The correlation coefficient between u and v will be given by

$$r_{uv} = \frac{s_y^2 - s_x^2}{\sqrt{(s_y^2 - s_x^2)^2 + 4\, s_x s_y \, cosec^2 2\alpha}}$$

if x and y are uncorrelated and their means are zero.

Ans: (i) Since cov (u, v) = 0

$$(cos^2\alpha - sin^2\alpha)\, cov(x, y) - sin\alpha . cos\alpha \left(s_x^2 - s_y^2\right) = 0$$

$$\Longrightarrow \cos 2\alpha . rs_x s_y = \frac{1}{2}\sin 2\alpha \left(s_x^2 - s_y^2\right) \Longrightarrow \tan 2\alpha = \frac{2rs_x s_y}{s_x^2 - s_y^2}$$

(ii) cov (u, v) = $\left(s_y^2 - s_x^2\right)\sin\alpha \,.\cos\alpha = \frac{1}{2}\left(s_y^2 - s_x^2\right)\sin 2\alpha$

$$s_u^2 = s_x^2 cos^2\alpha + s_y^2 sin^2\alpha$$

$$s_v^2 = s_y^2 cos^2\alpha + s_x^2 sin^2\alpha$$

$$s_u^2 \,.\, s_v^2 = s_x^2 s_y^2 \,(cos^4\alpha + sin^4\alpha) + (s_x^4 + s_y^4\,) sin^2\alpha cos^2\alpha$$

$$= s_x^2 \,.\, s_y^2 \,(\,1 - 2sin^2\alpha \,.\, cos^2\,\alpha\,) + \left(s_x^4 + s_y^4\,\right) sin^2\alpha . cos^2\alpha$$

$$= sin^2\alpha . cos^2\alpha \left(\, s_x^4 + s_y^4 - 2s_x^2 s_y^2\right) + s_x^2 s_y^2$$

$$= \frac{1}{4}\, sin^2 2\alpha \,(s_y^2 - \, s_x^2)^2 + \, s_x^2 s_y^2$$

So $r_{uv} = \dfrac{\frac{1}{2}\left(s_y^2 - s_x^2\,\right) \sin\, 2\alpha}{\sqrt{\frac{1}{4} sin^2 2\alpha \,(s_y^2 - s_x^2\,)^2 + s_x^2\, s_y^2}}$

$$= \frac{\left(s_y^2 - s_x^2\,\right)}{\sqrt{(s_y^2 - s_x^2\,)^2 + \, 4 \, s_x^2\, s_y^2 \, cosec^{\,2} 2\alpha}}$$

Q8. Given u = cx + dy and v= cx – dy, and r is the correlation coefficient between x and y. If u and v are uncorrelated , prove that

$s_u s_v = 2cds_x s_y \sqrt{1 - r^2}$

Ans: $cov(\,u, v) = 0 \implies c^2 s_x^2 - \, d^2 s_y^2 = 0 \implies cs_x = ds_y$ ……(1)

Now $s_u^2 = c^2\, s_x^2 + \, d^2 s_y^2 + \, 2cd\, s_x s_y r$

$= 2c^2\, s_x^2 + 2\, c^2 s_x^2 r = 2c^2\, s_x^2\, (1 + r)$ ……(2)

$s_v^2 = c^2\, s_x^2 + \, d^2 s_y^2 - \, 2cd\, s_x s_y r = \, 2c^2\, s_x^2 - 2\, c^2 s_x^2 r = 2c^2\, s_x^2\, (1 - r)$

……..(3)

(2)X(3) $\implies s_u^2 \,.\, s_v^2 = 4c^4\, s_x^4\, (1 - r^2)$

From (1), $s_u^2 \,.\, s_v^2 = 4c^2 d^2\, s_x^2\, s_y^2\, (1 - r^2) \implies s_u s_v = 2cds_x s_y \sqrt{1 - r^2}$

Q9. For 5 pairs of values of x and y, the values of x+y are 24, 28, 30, 33,

35 and variances of x and y are 6 and 2 respectively. Calculate the correlation coefficient between x and y.

Ans: Take u= x – 30

u: -6, -2, 0, 3, 5

$\sum u = 0 \; and \; \sum u^2 = 74$

$s_u^2 = \frac{74}{5} - (\frac{0}{5})^2 = 14.8$

Since variance has no effect on the change of origin $s_u^2 = s_z^2$

Now $s_z^2 = s_x^2 + s_y^2 + 2\, s_x s_y r_{xy} \implies 14.8 = 6 + 2 + 2\sqrt{12}\, .\, r_{xy}$

$\implies r_{xy} = 0.98$

Prepared by

Sanjay Bhattacharya